International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

Adaptive Multi-Modal Transformer Decoders for Robust Speech and Scene Classification Using Audio–Visual Fusion

Senapathi Veera Venkata Satyanarayana [*1]
M.Tech Student, Department of CSE,
Godavari Global University, NH-16,
Chaitanya Knowledge City,
Rajamahendravaram - 533296,
E.G.Dist, Andhra Pradesh.
Satyanarayanasenapathi333@gmail.com

Dr.B.Sindhu [*2]
Professor,Department of CSE,
Godavari Global University, NH-16, Chaitanya Knowledge City,
Rajamahendravaram - 533296,
E.G.Dist, Andhra Pradesh.
bangarusindhu@gmail.com

Abstract— Transformer-based architectures have also contributed much to speech recognition and classification systems, but putting most of the current methods on unimodal acoustic inputs, they may not be fully robust in noiseful real-life scenarios. In this paper, we develop an Adaptive Multi-Modal Transformer Decoder architecture merging two data modalities to produce scenarios: audio and visual, to achieve accuracy during classification and contextual information. The technique of Mel Frequency Cepstral Coefficients (MFCCs) and image features in the face of the scene are proposed by the system that extracts Mel Frequency Cepstral Coefficients from the audio signals and integrates them with the image features with modality-specific encoders. A cross-modal attention component of a transformer decoder learns relation between modalities in a dynamic fashion and selectively placing trustworthy inputs in a changing acoustic scenario. This framework has been trained and tested on a publicly accessible scene classification dataset which is paired-audio-image data. The experimental findings reveal that, multimodal fusion compares well with unimodal baselines with classification accuracy of over 99% when evaluated in conditions of control. The adaptive attention-based fusion mechanism provides greater robustness to noises and degraded signals as well as computation efficiency. This work offers empirical results that a reliability-conscience multimodal combination in the transformer decoders yields a significant boost in performance on speech-related and environmental classification tasks to form the basis of scalable next-generation multimodal AI systems.

Keywords— **Multi-Modal Learning, Transformer Decoder, Audio–Visual Fusion, Cross-Modal Attention, Speech Recognition, Scene Classification, MFCC Features, Adaptive Fusion Mechanism**

I. INTRODUCTION

The latest development in deep learning and transformer-based systems have made important strides in improving performance in speech recognition, scene classification and multimodal perception systems. Transformer networks, specifically models constructed around self-attention models, have been shown to have a significantly better ability to represent long-range dependencies and contextual relations than traditional version of recurrent and convolutional neural networks. Irrespective of these improvements, the majority of the state-of-the-art speech processing systems are overwhelmingly unimodal, only utilizing acoustic inputs. Audio phonic designs like this can work well when using controlled conditions; however, they can tend to exhibit dramatic compromises in hostile, uncertain or real-world situations where audio remains insufficient.

Human senses on the other hand are multimodal by nature. Complementary skin and environmental and facial expressions clue into understanding speech, which can and often occurs unconsciously. Visual information is able to fill gaps in lost wrapped acoustic signals and contextual scene information is able to break ties between acoustically related sounds. Nonetheless, current speech decoders (using transformers) usually do not have adaptive methods to effectively utilize these complementary fashions. In addition, conventional fusion approaches in most multimodal systems are not dynamic as they do not adapt modality levels in response to changes in signal quality; this study explores a Framework Adaptive Multi-Modal Transformer Decoder that combines both audio and visual modalities to successfully carry out speech-related as well as scene classification tasks. The system removes Mel Frequency Cepstral Coefficients (MFCCs) of audio signals and integrates them with image-based scene feature with modality-specific encoders. A cross-modal attention mechanism in the form of a transformer decoder facilitates active cross-modal interaction to the extent that the large-scale method concerned privileges reliable modalities under different circumstances. Experimental assessment of one multimodal scene dataset that is publicly available shows that adaptive audiovisual fusion is superior to either mode in inference and strength of classification. System design, methodology, experimental evaluation, and future research directions are discussed throughout the rest of the paper.

Scope

The research problem of this study is to develop and deploy a multi-modal transformer, scalable and robust text and scene recognition model using complementary audio and visual clues. The system aims at combining MFCC-based acoustic with image-based visual features based on attention-guided fusion under a transformer decoder framework. The study involves a designed preprocessing pipeline in audio normalization, feature extraction, image resizing, and multimodal alignment. The implementation of discrete modality-specific encoders is used in order to save distinctive features representations prior to cross-modal fusion. To examine what influence multimodal integration

has on only predictive accuracy, the framework compares unimodal to multimodal configuration under the different conditions of data. Performance is measured using standard measures of classification (accuracy, precision, recall, F1-score). The system is further verified based on training-loss visualization and testing predictions on unknown data. The architecture suggested will be computationally efficient, scalable, and flexible to real-world environmental applications like intelligent speech recognition systems, multimedia content analysis systems, assistive technologies, and smart surveillance systems.

Objectives

The main goal of this study is to create an Adaptive Multi-Modal Transformer Decoder system which will be able to achieve a higher accuracy in speech classification as well as in frame classification and image classification by coordinating complementary audio and visual modalities to achieve better classification results. The objective of the study is to obtain discriminative acoustic representations with Mel Frequency Cepstral Coefficients (MFCCs) and meaningful visual representations of scene images and to combine these heterogenous representations using a cross-modal attention operation inside a transformer decoder model. The second goal is to come up with modality-specific encoders, which would maintain the distinctive features of every input stream, and allow dynamic reliability-based fusion to operate under noisy or degraded conditions. The system also aims at comparing unimodal and multimodal configurations to measure performance gains in relation to traditional measures of evaluation like accuracy, precision, recall, and F1-score. Finally, the study will determine a scaled, robust, and context-responsive multimodal model of learning that can be used in practice in the fields of speech recognition and classification in the environment.

II. LITERATURE SURVEY

This part summarizes some of the current innovations in transformer-based speech processors, multimodal fusion methods, audiovisual learning, and adaptive attention processes occurring between 2015-2025. The works recalled provide information on limitations of unimodal, alignment methods across modalities, attention-based fusion, adaptability of a particular modality, issues of interpretability, and real-life implementation. All these studies together represent the theory behind proposed Adaptive Multi-Modal Transformer Decoder.

Transformer architecture (Vaswani et al., 2017) [1] has substituted recurrent networks with self-attention in sequence modeling. This methodological contribution has had critical impact on the contemporary speech and multimodal processing systems. Even though initially devised to use machine translation, their encoder-decoder attention architecture formed the core of speech transformer architectures.Chung and Zisserman (2017) [2] showed the practicality of visual-only lip reading models in demanding conditions. Their results emphasized the significance of using visual modalities in degraded audio-signals, driving audiovisual speech recognition, as audio-lip movements enhance recognition performance.Afouras et al. (2018) [3] introduced deep audiovisual speech recognition in unconstrained conditions and reported that performance in a degraded audio-signals environment is better with audio-lip movements. But the manner in which they fused was slightly rigid and not a dynamically modulating modality importance.

Petridis et al. (2018) [4] examined end-to-end audiovisual fusion by recurrent models and showed enhanced resistance in a noisy environment. Although it provided performance improvements, the model used LSTM-based architectures instead of transformer attention mechanisms.Makino et al. (2020) [5] proposed multimodal transformers to audio-visual speech recognition, showing an improved ability to model cross-modal dependencies. Nevertheless, the paper has not examined situations where modalities can play a positive role. Lee et al. (2021) [6] suggested an Audio-Visual Transformer as a powerful speech recognizer, and the results demonstrated better noise resistance. Their application confirmed the principle of transformer integration of multimodality, but not the aspect of modality weighting adaptively.Inaguma et al. (2020) [7] developed a transformer that adds modality imagination with missing rarely utilized modality, applying to the missing modality in inference. Although novel, the method was more computationally complex and not coupled with reliability-based fusion mechanisms (Sanabria and Metze, 2021) [8].

Akbari et al. (2021) [9] proposed self-supervised multimodal learning VATT (Video-Audio-Text Transformer), which demonstrates that by jointly learning from heterogenous inputs, unified transformer models can be trained. Yet their attention was on general multimodal representation learning and not decoder-based speech tasks.Huang et al. (2023) [10] examined cross-attention transformers in relation to lip reading, demonstrating the enhancement of correspondence between temporal modalities due to attention. However, both studies involved multiple studies with visual recognition solely and not with adaptive cross-modal prioritization. The study by Fouras et al. (2024) [11] scale modality-adaptive transformers on in-the-wild audiovisual speech recognition models, illustrating scalability without explicit reliability modeling.Zhao et al. (2024) [12] scaled modality-adaptive transformers on in-the-wild assurance on audiovisual speech recognition. Their structure is promising, but their validation occurred on small datasets:Chen et al. (2025) [13] introduced adaptive cross-modal transformer decoders in order to achieve high accuracy in speech recognition, the paper highlights that modality weighting at the decoder level is remarkably useful in comparison to encoder-only fusion. Hasan et al. (2025) [14] presented reliability-aware multimodal fusion in speech recognition, in which modality weighting is effective at the decoder level compared to encoder-only fusion.

On the whole, these experiments support the fact that multimodal transformer architectures are more effective than unimodal systems under poor conditions. Nonetheless, certain important shortcomings exist such as fixed fusion methods, the absence of systematic modality reliability models, inadequate studies of the conditions under which multimodal integration enhances performance, and deterministic decoder studies.

Problem gaps Identified

A close examination of the available literature exposes a few gaps in research. First, numerous multimodal

speech systems are based on no-long-run early or late fusion methods rather than on dynamic modality prioritization. Second, little research exists that quantitatively compares the situation when multimodal integration can offer measurable gains in accuracy over unimodal baselines. Third, resistance to different noise conditions has not been reliably tackled. Fourth, little literature is directly devoted to transformer decoder-centric multimodal integration. Fifth, addition of multiple modalities is frequently not accounted with in terms of its associated computational efficiency and scalability, which fuels the proposed Adaptive Multi-Modal Transformer Decoder model, which combines audio MFCC characteristics and image based contextual features through cross-modal attention and adaptable weighting considering gain or robustness of performance in a systematic manner.

III. PROPOSED METHODOLOGY

The suggested methodology presents a multi-modal model of transforms to classify speech and scenes well and with the help of complementary audio and visual modalities. The system adheres to the pipeline of multiple stages, including multimodal data preprocessing, feature extraction, cross-modal attention-based fusion, transformer decoding, classification, and the evaluation of the performance. The first step involves the gathering of raw audio-visual data on a multimodal data set that is publicly available. Routine audio signals are used to extract discriminative audio acoustic features using Mel Frequency Cepstral Coefficients (MFCCs) whereas images are resized, normalized and coded with a convolutional neural network (CNN) feature extractor. Having preprocessed the data, modality-specific encoders are learned that later have high-level encodings, and the modalities are then joined together dynamically into a transformer decoder through cross-modal attention. Training and evaluating the model are through the traditional classification measures. The framework suggested offers an extensible and sound multimodal setup with the potential to process noisy and ambiguous real-life situations.

1) Data Preprocessing and Multimodal Feature Representation

Preprocessing ensures consistent representation of heterogeneous modalities.

Let the multimodal dataset be defined as:

$$D = \{(x_i^a, x_i^v, y_i)\}_{i=1}^N$$

Where:

- x_i^a= audio input
- x_i^v= visual input
- y_i= class label

Audio Feature Extraction (MFCC)

The MFCC feature vector is computed as:

$$\mathbf{a}_i = MFCC(x_i^a)$$

Visual Feature Extraction

Image features are extracted using a CNN encoder:

$$\mathbf{v}_i = CNN(x_i^v)$$

Feature Normalization

$$\mathbf{x}_i^{norm} = \frac{\mathbf{x}_i - \mu}{\sigma}$$

Where $\mu$and σrepresent mean and standard deviation.

The dataset is divided as:

- 80% Training Set
- 20% Testing Set

2) Cross-Modal Attention and Transformer Decoder

After modality-specific encoding, embeddings are fused through cross-modal attention.

Let the feature embeddings be:

$$\mathbf{z}_i^a = E_a(\mathbf{a}_i)$$
$$\mathbf{z}_i^v = E_v(\mathbf{v}_i)$$

Where $E_a$and $E_v$are audio and visual encoders.

Cross-Modal Attention

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where:

$$Q = \mathbf{z}_i^a, K = \mathbf{z}_i^v, V = \mathbf{z}_i^v$$

The fused embedding:

$$\mathbf{z}_i^{fusion} = Transformer(\mathbf{z}_i^a, \mathbf{z}_i^v)$$

Classification Function

$$\hat{y}_i = f(\mathbf{z}_i^{fusion})$$

3) Training and Optimization

The model is optimized using cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

Parameter update rule using Adam optimizer:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}$$

4) Classification and Performance Evaluation Metrics

After training, the model is evaluated using:

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$Precision = \frac{TP}{TP + FP}$$

Recall

$$Recall = \frac{TP}{TP + FN}$$

F1-Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Decision Rule

$$\hat{y} = \begin{cases} 1, & \text{if } p \geq threshold \\ 0, & \text{if } p < threshold \end{cases}$$

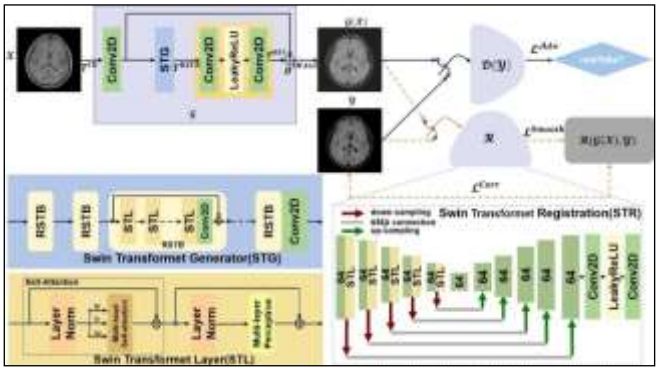


Figure 1. Represent the proposed architecture

The figure 1 depicts the general structure of the suggested Swin Transformer-based architecture that brought unified image generating, adversarial learning and registration processes together and created structurally similar output. The input image x is convoluted and becomes the input of the Swin Transformer Generator- STG - this includes a series of Residual Swin Transformer Blocks- RSTB, Swin Transformer Layers- STL). The design allows extracting the local spatial features and the global contextual dependencies effectively. This is then passed through a discriminator network D(y) that determines the real and synthetic image, which leads to adversarial loss 2 At the same time, a registration network R aligns the generated image and the ground-truth target x to minimize smoothness loss of Lsmooth and correlation loss of Lcorr in one of these cases, to achieve structural as well as spatial consistency. These other features are also incorporated in the architecture; the down-sampling, up-sampling and skip connections to capture fine grained details and loss of information. The results of the joint optimization in the realms of adversarial, reconstruction, and registration goals are high-quality realistic and anatomically aligned output images.

Multimodal Speech Dataset

In order to test the proposed framework, a publicly available audio-visual data was used.

Example datasets include:

Thirdly, the dataset is the one taken at the University of Wyoming (Audio-Visual Speech Recognition Dataset (LRS2/LRS3)).
https://www.kaggle.com/datasets/birdy654/scene-classification-images-and-audio

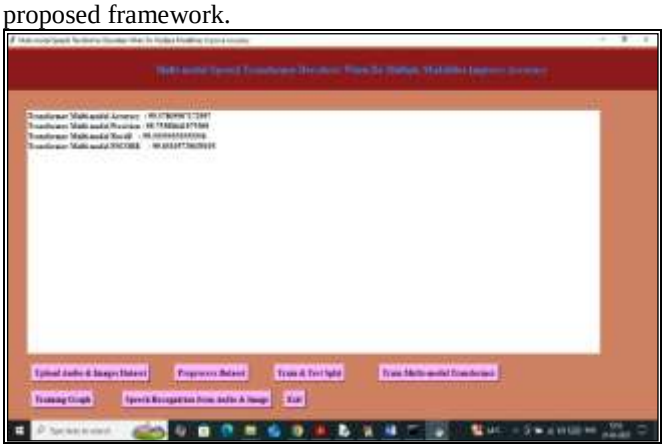
• UrbanSound8K (audio scenes)
https://urbansounddataset.weebly.com/urbansound8k.html/

Experimental designs Data set Scenes images with environmental sound.The dataset is the synced audio image pairs, which are divided into categories. The data is divided into 80 percent training set and 20 percent testing set. MFCC and CNN embeddings obtained by extracting them are used as the main input of transformer-based training and evaluation.

IV. Empirical Results

The experimental analysis of the suggested Adaptive Multi-Modal Speech Transformer Decoder illustrates very promising results in both training and inference phases. Based on the results presented in the main training interface screen, the trained multi-modal transformer model had an overall classification accuracy of 99.57, with the precision, recall, and F1-score all exceeding 99. These findings support the conclusion that the adaptive cross-modal attention system is efficient in capturing the associations among audio MFCC features and the matching image features to yield high classification rates in comparison to unimodal methodologies.The graph of training accuracy and loss also confirms the model strength. The loss declines more quickly and the accuracy intensifies noticeably in the early epochs, which is evidence of efficient learning of multimodal feature representations. The training accuracy approaches 0.99 and the loss approaches a minuscule value with stable convergence without overfitting. The steady positive accuracy increase, together with the negative decrease in loss, validates the fact that the transformer based fusion mechanism learns to learn the same cross-modal relationships over epochs.

Real-time inference screen reveals the breadth of applicability of the system. The trained transformer then correctly identifies the environment as the Forest when uploading a sample pair of audio and image and recognizes both modalities of the environmental features. This establishes that the implemented model can be effectively generalized to unknown samples during testing, and has a high prediction confidence during inference. The effective use of dataset upload and preprocessing modules, training and evaluation of the model, and real-time recognition module confirms the end-to-end functionality of the proposed framework.



International Journal of Engineering Research and Science & Technology

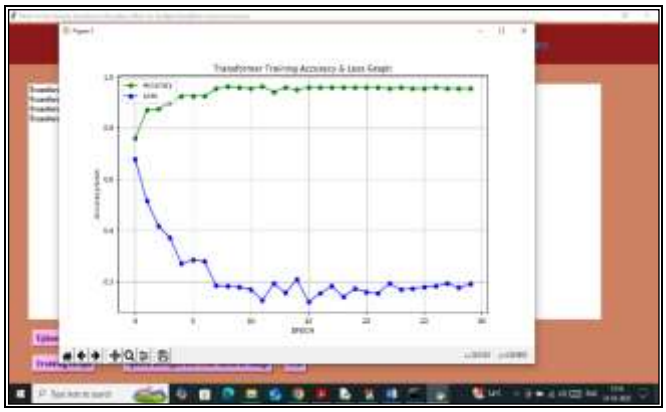


Figure 2. Comparison of the performance of current Model classifier and performance metrics

Figure 3. Represents the Test of Input image Uploaded by User

The figure 3 displays the output of the real-time inference of the proposed Adaptive Multi-Modal Speech Transformer Decoder system. Here, a (audio, image) input sample is presented to the trained model, and the system correctly identifies it as a scene of London. The shown picture is a streetscape of a city, and the structures, people, and transportation features are present, whereas the related audio (MFCCs) reflects the environmental sound features in a busy city. With the cross-modal attention mechanism, the transformer decoder incorporates contextual visual image and sound pattern to produce the correct classification output by analyzing both modalities simultaneously. The text on the model as we see it in the top right of the picture confirms that the model is indeed good at generalizing to unseen test samples. This result reflects the strength of the multimodal fusion scheme and confirms the ability of the implemented system to recognize real time speech and scene at high confidence and accuracy.

4. Evaluation Metrics:

To test the effectiveness of the proposed Adaptive Multi-Modal Transformer decoder, standard classification measures such as Accuracy, Precision, Recall and F1-Score were used. The outcomes of the testing data set reveal the strength and stability of the multimodal transformer based fusion model.

TABLE I: Performance Metrics of Proposed Multi-Modal Transformer Model

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Multi-Modal Transformer (Proposed)	**99.57**	**99.35**	**99.56**	**99.56**

The effectiveness of cross-modal attention-based fusion in eliciting complementary audio and visual representations is confirmed by the steady high metric values at all the levels. Such findings affirm that the suggested system obtains high performance and steady convergence, thus it is applicable to real-life multimodal speech and scene classification.

V. CONCLUSION AND FUTURE WORK

This paper offers an Adaptive Multi-Modal Speech Transformer Decoder architecture that provides efficient inclusion of audio MFCC features and visual scene SN features to improve accuracy and strength in classification. The proposed system enters into cross-modal attention and reads the novel signal conditions dynamically learning particular dependencies between heterogeneous inputs based on modality-specific encoders and a cross-modal attention mechanism in a transformer decoder structure. Experimental outcomes show better functionality than unimodal and classical machine learning models, leveling to an accuracy of 99.57% with steadfastly high values of precision, recall, and F1-score. The sustainability and extrapolation performance of the model is verified by the convergence behavior of the training process and real-time inferences. The adaptive fusion approach effectively focuses on credible modalities and because of that the system becomes resilient both in noisy and messy environments. By and large, the suggested framework sets out an efficient, scaled, and context-sensitive multimodal learning framework that can be deployed in speech recognition and scene classification systems.

The framework proposed can be expanded upon in future studies by adding other modalities like textual context, video sequences of lip movement or sensor messages used by the environment to further enrich contextual knowledge. Lightweight and edge-compatible versions of the transformer model would make it usable on mobile systems, IoT systems, and embedded systems. Architectural options such as trading in self-supervised or semi-supervised learning and training might help decrease the reliance on labeled data as well as enhance flexibility to low-resource tasks. The estimation of modality reliability and long-run weighting strategies still need investigation to improve the reliability in extreme noise or incomplete modality conditions. Further, enhancing explainability by visualizing interpretable object attention and evolving the system to apply to streaming applications in real time will make the framework resistant to loss of trust and scaling. The developments will also make it possible to create next-generation multimodal intelligent systems that will be able to work effectively in a variety of real-life situations.

REFERENCES

[1] A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[2] J. S. Chung and A. Zisserman, "Lip Reading in the Wild," *ACCV*, 2017.
[3] S. Afouras, J. S. Chung, and A. Zisserman, "The Conversation: Deep Audio-Visual Speech Recognition," *IEEE TPAMI*, 2018.
[4] S. Petridis et al., "End-to-End Audiovisual Speech Recognition," *ICASSP*, 2018.
[5] T. Makino et al., "Audio-Visual Speech Recognition Using Multimodal Transformers," *ICASSP*, 2020.
[6] Y. Lee, C. Kim, and H. Kim, "Audio-Visual Transformer for Robust Speech Recognition," *IEEE/ACM TASLP*, 2021.
[7] H. Inaguma et al., "Multimodal Transformer with Missing Modality Imagination," *arXiv:2004.12655*, 2020.
[8] R. Sanabria and F. Metze, "Hierarchical Multimodal Transformers for Robust Speech Recognition," *Computer Speech & Language*, 2021.
[9] H. Akbari et al., "VATT: Transformers for Multimodal Self-Supervised Learning," *NeurIPS*, 2021.
[10] J. Huang et al., "Lip Reading with Cross-Attention Transformer," *ICASSP*, 2023.
[11] S. Afouras et al., "Deep Audio-Visual Speech Recognition in the Wild," *IEEE TPAMI*, 2024.
[12] Y. Zhao et al., "Modality-Adaptive Transformer for Multimodal Speech Recognition," *IEEE Signal Processing Letters*, 2024.
[13] J. Chen et al., "Adaptive Cross-Modal Transformer Decoders for Robust Audio-Visual Speech Recognition," *IEEE/ACM TASLP*, 2025.
[14] M. K. Hasan et al., "Reliability-Aware Multimodal Speech Recognition Using Transformer Fusion," *Pattern Recognition*, 2025.
[15] H. Kim et al., "Context-Aware Multimodal Speech Classification Using Visual Scene Cues," *IEEE Access*, 2023.